Micro-Heterogeneity of Human Saliva Peptide P-C Characterized by High-Resolution Top-Down Fourier-Transform Mass Spectrometry

Frédéric Halgand,[a]* Vlad Zabrouskov,[b] Sara Bassilian,[a] Puneet Souda,[a] David T. Wong,[d] Joseph A. Loo,[c] Kym F. Faull,[a] and Julian P. Whitelegge[a]

[a] The Pasarow Mass Spectrometry Laboratory, NPI-Semel Institute for Neuroscience and Human Behavior, David Geffen School of Medicine, University of California-Los Angeles, Los Angeles, California, USA
[b] Thermo Fisher Corporation, San Jose, California, USA
[c] Department of Chemistry and Biochemistry and Department of Biological Chemistry, University of California-Los Angeles, Los Angeles, California, USA
[d] School of Dentistry, University of California-Los Angeles, Los Angeles, California, USA

Top-down proteomics characterizes protein primary structures with unprejudiced descriptions of expressed and processed gene products. Gene sequence polymorphisms, protein post-translational modifications, and gene sequence errors can all be identified using top-down proteomics. Saliva offers advantages for proteomic research because of availability and the noninvasiveness of collection and, for these reasons, is being used to search for disease biomarkers. The description of natural protein variants, and intra- and inter-individual polymorphisms, is necessary for a complete description of any proteome, and essential for the discovery of disease biomarkers. Here, we report a striking example of natural protein variants with the discovery by top-down proteomics of two new variants of Peptide P-C. Intact mass measurements, and collisionally activated-, infrared multiphoton-, and electron capture-dissociation, were used for characterization of the form predicted from the gene sequence with an average mass 4371 Da, a form postulated to result from a single nucleotide polymorphism of mass 4372 Da, and another form of mass 4370 Da postulated to arise from a novel protein sequence polymorphism. While the biological significance of such subtle variations in protein structure remains unclear, their importance cannot be assigned without their characterization, as is reported here for one of the major salivary proteins. (J Am Soc Mass Spectrom 2010, 21, 868–877) © 2010 American Society for Mass Spectrometry

Saliva is excreted by the parotid, submandibular, sublingual, and Von Ebner's exocrine glands to lubricate the mouth, initiate digestion, and contribute to the defense of the oral cavity by providing protection with antibacterial, antifungal, and antiviral activities [1, 4]. These activities are attributed to proteins and other small molecules excreted in saliva. With easy availability and a noninvasive collection process, saliva is a useful fluid to use in the search for disease biomarkers. Two approaches are being used to characterize the salivary proteome. The first relies on standard bottom-up proteomic approach with tandem mass spectrometry (MS/MS) of tryptic peptides to compile a catalogue of salivary proteins and of predictable post-translational modifications [2]. However, bottom-up proteomic strategies suffer from technical problems such as variable digestion yield, variable peptide extraction efficiency from gels, missed cleavages, poor recovery of some peptides during liquid chromatography, and poor ionization efficiency of some peptides, leading to variable and incomplete coverage of protein sequences. In addition, bottom-up experiments only recognize predicted post-translational modifications, exacerbating this problem.

The second approach being used to characterize the salivary proteome uses a combination of top-down, with liquid chromatography (LC) fractionation of intact proteins coupled with intact protein mass measurements, and bottom-up approaches, to more completely characterize salivary proteins [3]. Provided that proteins can be relatively well separated from each other, analysis of the intact proteins using top-down strategies provides an alternative means to fully characterize the diversity of the human salivary proteome. Different protein variants are usually defined by their own unique molecular weight. An intact mass tag (IMT) refers to the experimentally measured mass of a whole

Address reprint requests to Dr. Frédéric Halgand, Laboratoire de Bioénergétique et Ingénierie des Protéines, Equipe de protéomique fonctionnelle et dynamique, UPR 9036-CNRS, 31 Chemin Joseph Aiguier, 13420 Marseille Cedex, France. E-mail: fhalgand@ifr88.cnrs-mrs.fr

* Current address: Laboratoire de Bioénergétique et Ingénierie des Protéines, Equipe de protéomique fonctionnelle et dynamique, UPR 9036-CNRS, 31 Chemin Joseph Aiguier, 13420 Marseille Cedex, France.

1044-0305/10/$32.00
doi:10.1016/j.jasms.2010.01.026

Published online February 1, 2010
Received September 29, 2009
Revised January 27, 2010
Accepted January 27, 2010

protein, and it reflects the presence of all covalent modifications [4]. This method has been promoted by the work of Kelleher and coworkers [5] and is now becoming an attractive alternative to obtain detailed information on protein structure [6]. Consequently, comprehensive saliva proteome characterization must involve the use of both bottom-up and top-down approaches to define the different protein variants [3, 7]. The combination of these approaches will eventually provide an extensive catalogue of salivary proteins with their detailed molecular characterization.

However, the salivary proteome has some properties that make it difficult to fully characterize. First is the presence of protein superfamilies, such as the abundant proline-rich proteins (PRPs) and histatins in which many different protein products arise by differential post-translational processing of a limited number of gene products. This leads to a high degree of sequence homology amongst the various products, exacerbated in some cases by the presence of repetitive sequence units that require extensive sequence coverage to distinguish between them. Secondly, salivary proteins are often subjected to extensive proteolytic processing at both N- and C-termini before secretion. The resulting products may be too small to produce the number of tryptic fragments normally required for unequivocal bottom-up identification or, as in the case of the PRPs, bottom-up identification may be compromised because the juxtaposition of proline and basic residues renders trypsin ineffective. Third, single nucleotide polymorphisms, alternative splicing, and post-translational modifications all contribute to further increase the heterogeneity of human saliva proteins [8]. While the goal of cataloguing the presence of different salivary gene products is simply accomplished with bottom-up approaches, the goal of completing detailed molecular characterization of these products is more complex.

Castagnola and colleagues used liquid chromatography (LC) combined with online electrospray ionization mass spectrometry (ESI-MS) to characterize the more abundant lower molecular weight protein variants in various saliva preparations [9, 10]. By supplementing low-resolution intact mass measurements obtained on quadrupole instruments with MS/MS data after proteolysis, a good start was made toward documentation of the processing and modifications of several classes of salivary proteins (e.g., histatin and statherin), and their variability between individuals [3, 7, 11–13]. More recently this approach was successfully used by Messana et al. to elucidate saliva protein trafficking in concert with post-translational events such as phosphorylation, sulfation, and cleavage of propeptides [14].

Our goal is to apply top-down mass spectrometry experiments to characterize human salivary proteins in significantly greater detail. Here, we report on experiments that distinguish variants of Peptide P-C, an abundant human salivary protein that is derived from in vivo cleavage of salivary acidic proline-rich phosphoproteins 1 and 2 (PRP superfamily), and that is also characterized by the lack of trypsin cleavage site. Acidic PRPs are thought to be involved in calcium homeostasis in the oral cavity and act by preventing calcium phosphate precipitation that is necessary for maintaining tooth enamel integrity. Also, recent studies have demonstrated that Peptide P-C can modulate blood glucose levels after feeding by inducing insulin release and by restricting glucagon release [15, 16]. The Peptide P-C variants have a nominal mass in the 4370–4372 Da range with overlapping isotopomer distributions that were resolved only with a Fourier-transform ion cyclotron resonance (FTICR) mass spectrometer operating at 750,000 resolving power. Manual data interpretation was needed to assign the newly reported variants because they would have been otherwise overlooked by automated data processing, as their primary structure was not in the sequence databases.

Experimental

Chemicals

All solvents (HPLC grade and otherwise), buffers, and reagents [guanidine HCl, acetonitrile, anti-protease cocktail, trifluoroacetic acid (TFA)] used were purchased from Sigma Aldrich.

Sample Collection

Saliva samples were obtained from five donors who were recruited for the Human Salivary Proteome Project. Donors were in good health and exhibited normal salivary function. Parotid saliva secretions were harvested using a saliva collector [17] fitted with a sterile 100 μL pipette tip. Stimulation of the salivary glands was provided by repeated topical application of a mild solution of citric acid (2%) to the dorsal surface of the tongue. Care was taken to keep the acid solution away from the collection area. Collection volumes were 500–2000 μL/donor. The collected samples were centrifuged (2600 $\times$ g, 15 min, 4 °C), and re-centrifuged (20 min) if the supernatant was not clear to the naked eye. Supernatants were transferred to new containers to which was promptly added aprotinin (1 μL/mL saliva, 10 mg/mL), sodium orthovanadate (3 μL/mL saliva, 400 mM), and phenylmethyl sulfonyl fluoride (10 μL/mL saliva, 10 mg/mL) before storage at −80 °C.

Sample Fractionation by LCMS+

Individual parotid saliva samples (between 500 μL and 1 mL) were dried by centrifugal evaporation, redissolved in aqueous guanidine-HCL (6 M, 100 μL), centrifuged (10,000 $\times$ g, 5 min, room temperature), and processed by combined liquid chromatography-mass spectrometry with on-line fraction collection [18] in which the supernatant was injected onto a polymeric reversed phase column (Polymer Labs PLRP/S 5 μm, 300 Å, 2 $\times$ 150 mm, 40 °C) previously equilibrated in

95% buffer A, 5% buffer B (A, 0.1% TFA in water; B, 0.1% TFA in acetonitrile), and eluted (150 μL/min) with an increasing percentage of buffer B (min/% B; 0/5, 5/5, 10/20, 70/50, 90/90). The eluent was passed through a UV detector (280 nm) before a flow splitter with fused silica capillaries to transfer liquid to an Ionspray source (50 cm, ~50 μL/min) and the fraction collector (25 cm, ~100 μL/min). Fractions (1 min) were collected into microcentrifuge tubes and stored at −20 °C for further analysis.

The Ionspray source was connected to a triple quadrupole mass spectrometer (API III+; Applied Biosystems) tuned and calibrated as described [19], scanning from *m/z* 600 to 2300 (orifice voltage ramped with *m/z* from 6 to −120, 6 s/scan). Data were processed using MacSpec 3.3, Hypermass, and BioMultiview 1.3.1 software (Applied Biosystems).

High-Resolution Top-Down Mass Spectrometry and Tandem Mass Spectrometry

These experiments were performed as previously described [8] on a 7 Tesla hybrid linear ion trap-FTICR mass spectrometer (LTQ-FT Ultra, Thermo Fisher Corporation, San Jose, CA, USA) fitted with an off-line nanospray source. Pooled fractions from the LCMS+ experiments, eluting between 19 and 21 min and encompassing all the detectable variants of Peptide P-C, were individually loaded into 2 μm i.d. externally-coated nanospray emitters (New Objective Inc., Woburn, MA, USA) and directly desorbed/ionized using a spray voltage of between 1.2 and 1.4 kV (versus the inlet of the mass spectrometer). These conditions produced a flow rate of 20–50 nL/min. Ion transmission into the linear trap and further to the FTICR cell was automatically controlled to a 2×10^6 ion count target for both the full scan- and MS^2-FTICR experiments. The *m/z* resolving power of the FTICR mass analyzer was set to either 100,000 or 750,000 (defined by $m/\Delta m_{50\%}$ at *m/z* 400). Individual charge states of the multiply protonated protein molecular ions were selected for isolation and collisional activation in the linear ion trap, followed by the detection of the resulting fragments in the FTICR cell. For FTICR-MS/MS experiments, parameters were chosen to fragment the full isotopic mass of the +5 charge ion with the view to increase detection of product ions while checking for homogeneity of the selected peak. For the collisionally activated dissociation (CAD) studies, the precursor ions were activated using 12% to 15% normalized collision energy at the default activation *q*-value of 0.25. Additional studies were conducted in which the precursor ions were guided to the FTICR cell and further fragmented using electron capture (ECD) using the following instrument settings (5% to 10% normalized collision energy, 50 ms delay, and 10 ms duration) and infrared multiphoton (IRMPD, instrument settings of 50% normalized collision energy, 50 ms delay and 20 ms duration) dissociation experiments. In both cases, the fragmentation efficiency was optimized to maximize product ion signal intensity. Ion count target was the same for CAD, IRMPD, and ECD MS/MS experiments with a value of 2×10^6.

FTICR Data Analysis

FTICR spectra, from an average of 50–500 transient signals, were examined with a combination of manual and automatic procedures. Monoisotopic mass lists ($s/n = 1.1$, fit 0%, remainder 0%, averagine table set to averagine) were prepared using XtractAll (Xcalibur 2.0, Thermo Fisher, Bremen, Germany). ProSight PTM (https://prosightptm.scs.uiuc.edu) and ProSight PC (Thermo Fisher) software suites were used with a threshold of 15 ppm and the deltamass feature deactivated, with custom post-translational modifications as required. Interpretation was a manual, iterative process as different sequences and post-translational modifications were independently tested to maximize the number of product ions matched. Nomenclature for assignment of peptide/protein ions was according to Roepstorff and Fohlman [20].

P-Score values reflect the match of the proposed primary structure with the peaklist data; the lower the score, the higher the confidence in the proposed sequence [21]. We also used a manual P-Score that similarly reflects the confidence in data interpretation, but relies on masses of product ions that matched the sequence, updated with product ions manually identified in the MS/MS spectra. Manual P-Scores were calculated in ProSight PTM using the Manual Single Protein Mode. Extracted peak lists are provided in the Supplemental Materials, which can be found in the electronic version of this article.

Putative Deamidation of Peptide P-C

To determine if Peptide P-C was prone to a time-dependent deamidation process at residue Q14, an HPLC fraction containing this material was split into two equal parts. One part (~20% acetonitrile, 0.1% TFA) was untreated, and the other dried in a vacuum centrifuge and re-dissolved in ammonium bicarbonate (pH 7.5). Both parts were incubated for 10 days at 37 °C. Each day, an aliquot from each was analyzed by recording the MS and CAD MS/MS spectra, and the isotopic profiles of the b_{15}-product ions were compared.

Results

Peptide P-C is derived from a salivary PRP (PRH1 and allele PRH2, accession number P02810). PRH1 is easily identified in bottom-up studies, although these datasets cannot distinguish between the various products of PRH1 that accumulate in saliva. Peptide P-C is a peptide with a mass of ~ 4371 Da—outside the mass range usually employed in bottom-up protocols. It is easily

detected as an IMT in a reversed-phase chromatogram. Based on the preliminary analyses, a number of factors suggested micro-heterogeneity within the Peptide P-C population. These factors included a bimodal elution profile and a variable charge-state distribution across this chromatographic peak (data not shown). Consequently, fractions collected concomitant with detection of the 4371-Da IMT during LC-MS+ were subjected to further analysis by top-down MS.

High-resolution (R = 100 000 @ *m/z* 400) intact mass measurements and the IRMPD MS/MS spectra clearly confirmed the presence of the known Peptide P-C sequence with high confidence when the data were processed automatically with Xtract and ProSight PTM (Figure 1c and Table 1 of Supplementary Material). However, by manually comparing experimental and theoretical isotopomer profiles, the spectra indicated the presence of both lighter and heavier forms than that predicted from the known Peptide P-C sequence (P02810). For example, examination of the pentuply-charged precursor ion showed a minor isotopomer 1 Da lighter than the monoisotopic peak (Figure 1a). Furthermore, the isotopomer cluster of signals assigned as a mixture of doubly-charged b22-NH_3 and b21-H_2O IRMPD product ions similarly displayed the presence of a peak 1 Da lighter than the monoisotopic peak (Figure 1b). This ion at *m/z* 1066.0447 was later assigned to the b21-H_2O product ion of the lighter form of Peptide P-C (Variant 1). At the same time, the isotopomer cluster for the b15 IRMPD product ion presented an overabundance of the first ^{13}C component, indicating the presence of a species 1 Da heavier than the monoisotopic peak (data not shown). These features led to further investigations of the possible presence of two variants of the peptide that differ by the nominal mass of ± 1Da.

Additional MS and MS/MS experiments were performed at ultra-high-resolution (R = 750,000 at *m/z* 400) in an attempt to better resolve the microheterogeneity. The isotopomer distribution of the molecular ion signals for the 5+-charged ion of Peptide P-C at *m/z* 875.0459 revealed a mixture of three putative variants (Figure 2). These forms have measured monoisotopic masses of 4367.2570 (Variant 1), 4368.1855 (consistent with residues 123-166 of accession number P02810), and 4369.1696 (Variant 2) Da (Table 1).

To account for the two extra variants, we first proposed the heavier form as due to N-to-D or Q-to-E SNP, or the possible occurrence of a single deamidation. D-to-N SNPs have been reported for positions 20 and 66 of PRH1 (Peptide P-C is residues 123-166 of PRH1; GeneID: 5554) [22, 23]. Although a Q-to-K variant has also been reported at position 163 within the Peptide P-C sequence [23], this was considered unlikely in this sample because such a change was inconsistent with the mass measurement accuracy available on the FTMS instrument. The lighter form, Variant 1 (Table 1), has a significant positive mass defect that is hypothesized to result from a sequence change eliminating oxygen atoms. The explanation for this lighter species was not immediately apparent, and therefore the deduced protein sequences from the three calculated open reading frames of full-length PRH1/2 mRNA sequence (P02810) containing the Peptide P-C sequence (NM_005042) were

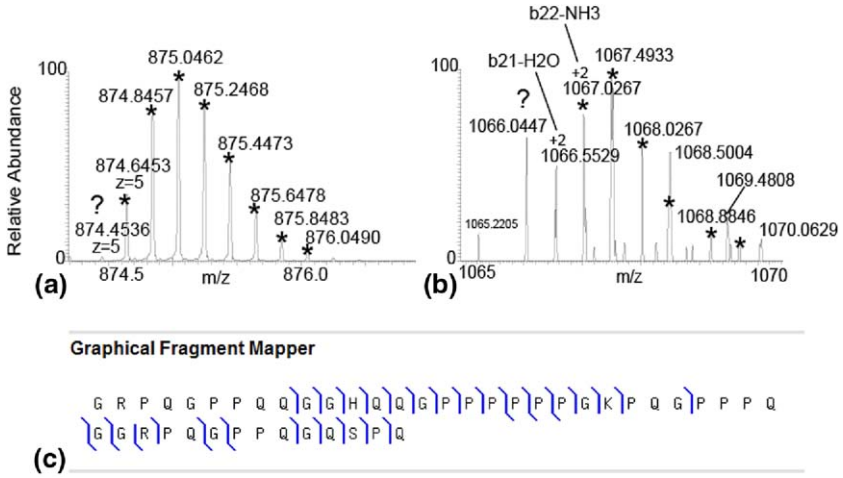

Figure 1. (**a**) Expanded view of the 5+-charged precursor ion at *m/z* 875.05 recorded at a resolution of 48,000 before LTQFT IRMPD MS/MS fragmentation. The asterisks show the calculated intensity of the predicted isotopomer signals based upon the elemental composition for Peptide P-C (C_{189} H_{290} N_{64} O_{57}). A small apparent isotopomer 1 Da lighter than the monoisotopic peak is labeled with a question mark. (**b**) Expanded view of the isotopomer cluster of the b22-NH_3 and b22-H_2O product ions produced during IRMPD fragmentation of the precursor shown in (**a**). The asterisks show the calculated intensity of the predicted isotopomer signals based upon the elemental composition for the b22-NH_3 ion (C_{94} H_{138} N_{31} O_{27}). This spectrum reveals an additional isotopomer 1 Da lighter, labeled with the question mark. (**c**) Sequence coverage of *b*- and **y**-ions observed during IRMPD of the 5+ precursor ion at *m/z* 875.05 matched to the published sequence of Peptide P-C (P02810) using Xtract and ProSight PTM software.

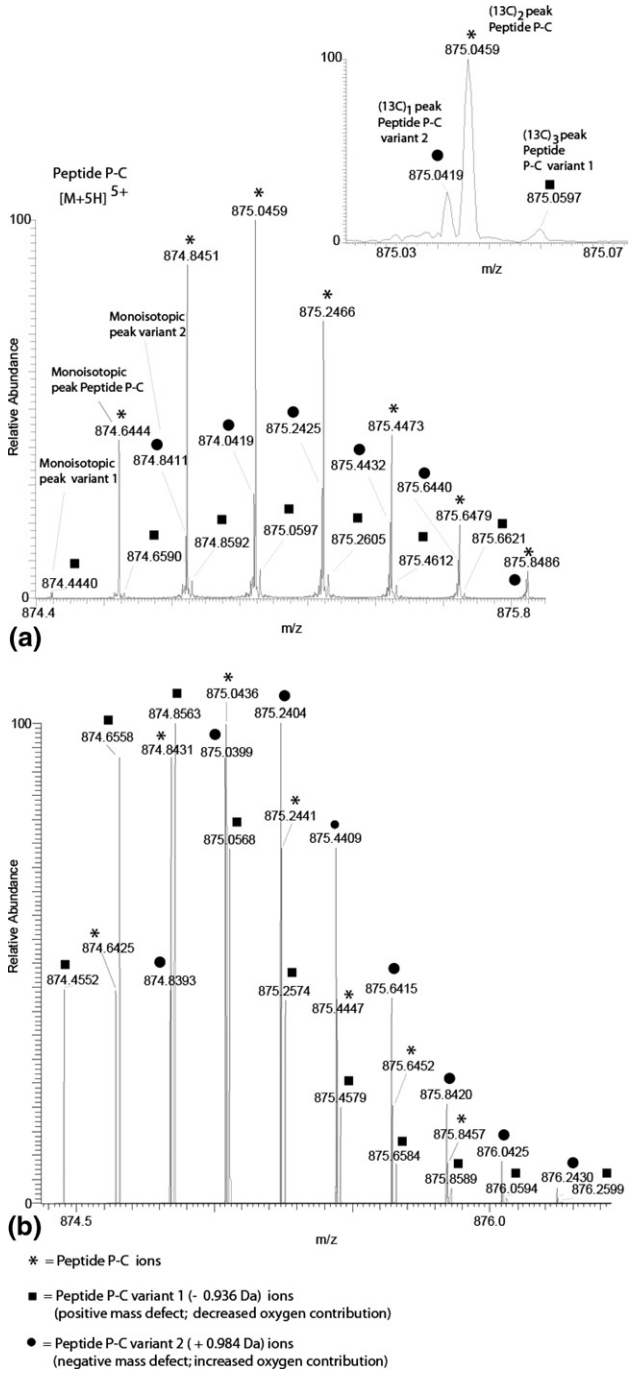

Figure 2. (**a**) Mass spectrum of the 5+-charged ion of Peptide P-C at ultra-high-resolution (R = 750,000 at *m/z* ~400). The three proposed Peptide P-C species (see Table 1), including the lighter Variant 1 postulated as replacement of QQGPP by PRPPR (−1 Da), the regular sequence of Peptide P-C, and the heavier Variant 2 postulated to arise from deamidation or a SNP (+ 1 Da), are respectively labeled (filled square), (asterisk), and (filled circle). The assignments are chosen to best account for the observed masses and mass defects. Experimental monoisotopic molecular weights deduced from the monoisotopic peak of each species are 4367.2570 Da (theoretical value: 4367.2386 Da, mass error = 0.0184 Da, 4.2 ppm) for Variant 1, 4368.1855 Da (theoretical value: 4368.1750 Da, mass error = 0.0105 Da, 2.4 ppm) for the regular form, and 4369.1696 (theoretical value: 4369.1590 Da, mass error = 0.0106 Da, 2.4 ppm) for Variant 2, respectively (see Table 1). (**b**) Theoretical isotopomer profiles of the three proposed variants overlaid, each normalized to 100% intensity.

examined to uncover sources of potential polymorphisms. Predictions of ORFs and their translations were performed with the Wise2 software from EMBL-EBI (http://www.ebi.ac.uk/Wise2/), and results obtained for the C-terminus of the protein corresponding to the Peptide P-C sequence are displayed in Figure 3. This revealed the presence of a homologous 16 amino acid in-frame sequence of Peptide P-C, immediately upstream (residues 106-122) of the regular Peptide P-C sequence (residues 123-140 of 123-166), which only differs by a short stretch of amino acids with the replacement of the regular sequence QQGPPP by PRPPR. This sequence replacement results in a calculated molecular weight for the Peptide P-C variant of 4367.2386 Da, or 0.9364 Da lighter than the mass calculated for the regular sequence. This is hypothesized to explain the presence of Variant 1. Though the genetic basis of this variant form is obscure at this point, it is somewhat consistent with previous literature reporting saliva protein genetic polymorphism for PRPs. Table 1 summarizes the three observed forms and their assigned sequences.

The veracity of these assignments was tested with additional MS/MS experiments specifically to locate *unique* product ions to confirm these sequence assignments. A unique ion is an unambiguously assigned product ion that cannot be explained by internal fragmentation, or loss of water, ammonia etc. Manual analysis of the IRMPD MS/MS spectrum from the 5+-charged precursor cluster at 750,000 resolution showed some product ions lacking shoulders, such as product ion y11 (*m/z* 1120.54), some with a single satellite peak such as the b41 product ion (*m/z* 1011.01, z = 4), and some with two additional peaks such as b42-NH_3 product ion (*m/z* 1028.51, z = 4, Figure 4). The fact that some product ions showed microheterogeneity while others did not is consistent with the true peptide microheterogeneity as concluded from Figure 2, and assuaged concerns that the results were some type of artefact of the FTICR experiment. However, no unique ions were found in either the CAD (data not shown) or IRMPD datasets that allowed unequivocal assignment of the location of the variant masses. Consequently, the origin of the Peptide P-C sequence microheterogeneity was sought from ECD experiments.

The analysis of ECD spectra showed that nearly all permitted cleavages were observed for the regular and Variant 2 forms (Tables 2 and 3 of Supplementary Material, and Figure 5). Diagnostic ions used to assess the presence of both variants are $z^{\bullet}$-product ions. Attention was paid to the presence of z′-product ions when the heavier form of Peptide P-C was searched. In that case, the experimental mass accuracy was sufficient to distinguish between z′- and $z^{\bullet}$-product ions from regular and heavier forms, respectively. The ECD experiments unequivocally confirmed the regular sequence and the presence of deamidation/SNP at position 14 (Gln) with a P-score of $6 \times 10e^{-79}$ for Variant 2 (Figure 5, and Table 3 of Supplementary Material). In

Table 1. Sequence of Peptide P-C variants and their masses

Protein ID/peptide P-C	Measured monoisotopic mass (Da)	Acc numbers (SwissProt)	Proposed sequences	Calculated monoisotopic MW (Da)	Mass accuracy (PPM)	Observations
Variant 1	4367.2570 Da		GRPQGPPQQGGHPRPPRPPPGKPQ GPPPQGGRPQGPPQGQSPQ	4367.2391	4.2	Replacement of QQGPPP by PRPPR
Regular	4368.1855 Da	P02810 123-166	GRPQGPPQQGGHQQGPPPPPPGKP QGPPPQGGRPQGPPQGQSPQ	4368.1755	2.4	Regular sequence
Variant 2	4369.1696 Da		GRPQGPPQQGGHQEGPPPPPPGKP QGPPPQGGRPQGPPQGQSPQ	4369.1595	2.4	SNP's at Q14

this particular experiment, the RMS was higher with a value of 7.14 ppm compared to the other searches. This is attributed to misassignment of monoisotopic versus ^{13}C peaks for the regular and Variant 2 (deamidated/SNP) forms as a result of incomplete peak lists for the minor sequence variant.

To distinguish between the occurrence of a deamidation process and a SNP, the isotopic profile of the b15 product ion that was used to identify the heavier form was monitored over several days. Peptide P-C was placed at 37 °C directly from the HPLC fraction (~20% acetonitrile, acidic pH) or re-dissolved in ammonium bicarbonate (pH 7.5) after drying. CAD MS/MS spectra recorded over the following 10 days revealed that the relative intensity of the ^{13}C peak was unchanged, allowing conclusion that the heavier form likely corresponds to a SNP rather than a deamidation process. In addition, the likelihood of such deamidation process occurring during sample handling/processing was deemed relatively low because the samples were not exposed to basic pH [24, 25].

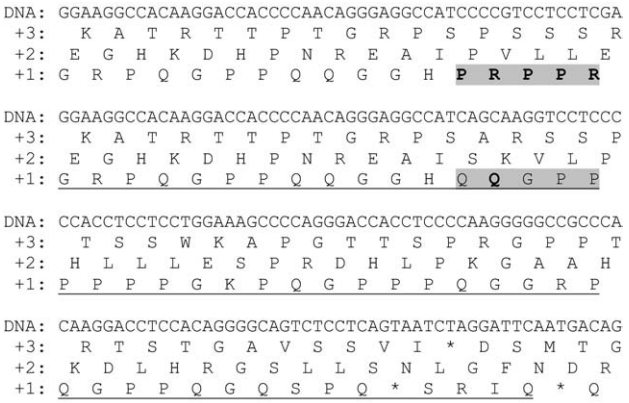

Figure 3. Calculated open reading frames (named +1, +2, and +3) from the full DNA sequence (NM_005042.2) of the proline rich protein (P02810) Residues 106-173 is shown) containing the Peptide P-C sequence (Peptide P-C sequence is residues 123-166, underlined sequence). Sequence alignment shows the tandem sequence repeats and reveals the possibilities for sequence recombination due to DNA slippage or alternative splicing. Taking into account these possibilities, it was postulated that the lighter Variant 1 could arise by replacement of residues 123-140 with 106-122, altering QQGPPP to PRPPR (shaded). Glutamine residue in bold in Q**Q**GPPP sequence corresponds to the amino acid that was found to be replaced by E in Variant 2.

However, we were unable to completely resolve the sequence of Variant 1. Although many predicted cleavages were observed by ECD, unique diagnostic product ions consistent with the hypothesized replacement of the QQGPPP sequence by PRPPR (Table 4 of Supplementary Material) could not be assigned from the datasets because there were no fragments produced within the putative PRPPR substitution sequence. Further experiments on samples enriched in this variant will be necessary to fully resolve the sequence and unambiguously assign the details of the mass change.

Discussion

High-resolution top-down MS based approaches to proteomics were first described by McLafferty and coworkers [26] and have since become a powerful method to embrace the full complexity of protein structure [27]. Following this demonstration, software was developed with a view to automate data analysis. This was achieved with development of Thrash [28] and ProSight PTM software [21]. Many of the most abundant proteins in human saliva are proteolytic cleavage products derived from larger precursors, and many of these feature proline-rich sequence repeats. While the physiologic significance of these sequences remains an interesting question, it is noteworthy that bottom-up proteomics strategies can report accurately on those gene products that are present, but poorly on the exact nature of the processed gene products. Analysis of the intact proteins using top-down strategies provides a means to accurately characterize the full diversity of the human salivary proteome. This study of Peptide P-C, a small ~4371 Da peptide abundant in human saliva, illustrates a difficult problem in top-down analysis.

In our article, we wondered if the term "top-down" or "middle-down" was the most appropriate to describe our work. Referring to the papers where the "top-down" term was described [26, 29], the most rigorous definition of a top-down experiment involves high-resolution measurement of an intact molecular weight value and direct fragmentation of protein ions in the gas phase. A new variant method, known as "middle-down" was recently introduced, which combines the benefits of limited proteolysis to generate large protein fragments that are then fragmented by tandem MS [30].

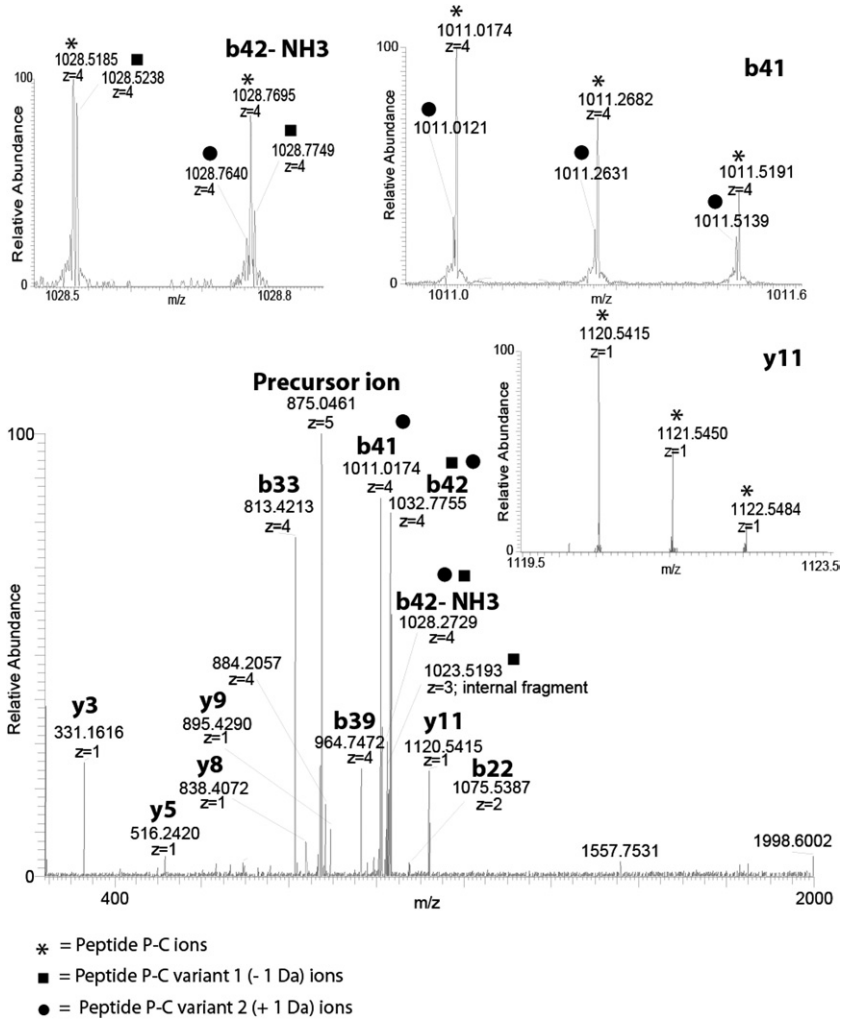

Figure 4. Ultra-high-resolution IRMPD spectrum of products arising from the 5+-charged parent ion of Peptide P-C (R = 750 000). Inserts show product ions that contain no shoulders such as the y11 ion at *m/z* 1120.54, a single satellite peak such as the b41 ion (*m/z* 1011.01, z = 4), or two additional satellite peaks such as the b42-NH_3 ion (*m/z* 1028.51, z = 4). Peaks were attributed to Variant 1 [− 0.936 Da, (filled square)], the regular Peptide P-C sequence (asterisk), and Variant 2 [+0.984 Da, (filled circle)]. Tables of assigned product ions for the three forms with automated and manually extracted peak list are shown in Supplementary Materials.

It appears to us that experiments we developed in our study lie somewhere in between "top-down" and "middle-down." However, our 4–5 kDa polypeptides are processing products of larger proteins that are endogenous and functional to the biological system, i.e., proteases were not added externally for the purposes of deriving amino acid sequence, as prescribed for a "bottom-up" or "middle-down" protocols. The 4–5 kDa size range is probably a reasonable lower limit for the "top-down" term because it is not routine to determine complete sequence information for a 40–50 amino acid polypeptide from a single MS/MS experiment (even with ECD/ETD). For all these reasons, we considered that our experiments refer to a top-down approach that would, in that case, embrace all proteins with a mass of 4–5 kDa and above that have not been further processed for the purpose of the study.

The original suspicion of the presence of Peptide P-C variants was the result of partial chromatographic separation and non-overlapping ESI charge state distributions recorded with a low-resolution mass spectrometer. However, full characterization of each variant requires identification of unique product ions that unambiguously solve the primary structure. In the case of PRPs, and Peptide P-C in particular, repeated use of the same amino acid residues and the presence of sequence repeats hamper the identification of unique product ions. Furthermore, the production of internal product ions under CAD or IRMPD conditions increases assignment ambiguity, again lowering the chances of identifying unique product ions. High-resolution and mass accuracy helps to limit doubt, but still many ambiguities remain because of identical chemical formulae for potential assignments of fragment ions. The lack of internal fragmentation in ECD experiments is an advan-

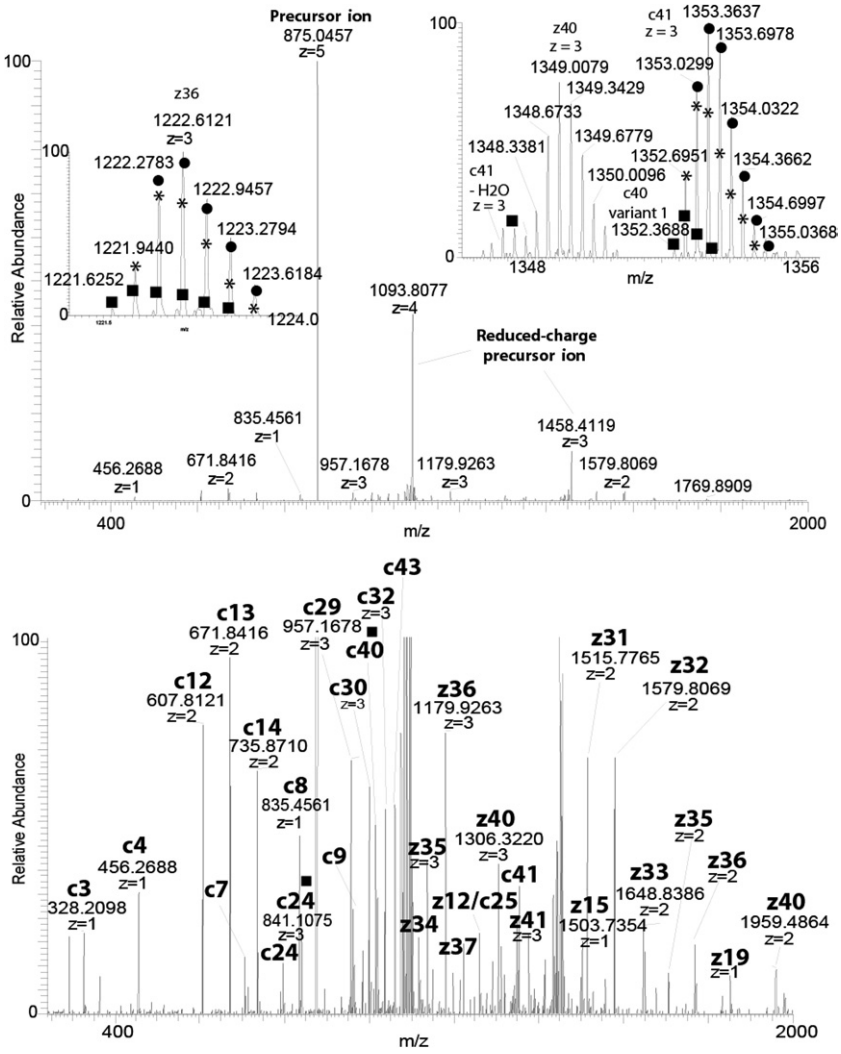

Figure 5. ECD MS/MS spectrum of the 5+-charged parent ion of Peptide P-C (R = 100 000). The upper panel shows the full spectrum revealing the additional lower charge states of the parent resulting from charge neutralization. The lower panel is the same spectrum with the ordinate expanded by 32-fold, showing the low-abundant fragment ions of interest. Inserts on the upper panel show expanded views of z36 and z40 fragment ions where detection of the monoisotopic peaks revealed the presence of all three variants. Peaks were attributed to Variant 1 [−0.936 Da, (filled square)], the regular Peptide P-C sequence (asterisk), and Variant 2 [+0.984 Da, (filled circle)]. However, there were no fragment ions in the ECD spectrum that unambiguously verified the presence of the PRPPR sequence that is postulated for Variant 1.

tage. However, in this work only the normal form and Variant 2 variant could be clearly defined, even though the experiments were done at ultra-high-resolution (R = 750,000) and with CAD, IRMPD, and ECD fragmentation.

When Variant 2 was investigated, the apparent mass accuracy of the product ions was lower than expected. This is assigned this to the fact that the Xtract software could not all three distributions, and thus incorrect peaks (e.g., ^{13}C of the regular Peptide P-C instead of the monoisotopic peak of the deamidated form) were used for assignments by the ProSight software. A solution to this problem lies in the modeling of theoretical profiles to better match data. ProSight was found to be reliable in the manual mode search but is completely reliant on the peak list from Xtract, dictating the need for additional manual processing. Solutions to this problem will require more highly developed algorithms that attempt to *interpret* the entire mass spectrum with capabilities for de novo sequencing, and not just matching to a database. Ideally, software for automated assignment of top-down data needs to consider all possible primary structure permutations at each residue, with some type of unbiased iterative process to refine the best match while taking into account the experimental isotopic variations.

Since 1971, there have been numerous reports of the presence of genetic polymorphism in salivary proteins [31]. Genetic recombination of PRPs leading to a variety of molecular weight variants were linked to the expression of different alleles [22, 23, 32, 33]. In a more specific

manner, Lyons et al. [34] reported the presence of insertion/deletion polymorphism within the PRP multigene family. Variant lengths were shown to coexist and the result of different numbers of tandem repeats in the third exons of several individuals. Homologous but unequal intragenic crossovers seemed to be a general phenomenon for the PRP genes. In contrast, PRH genes (Peptide P-C genes) were shown to contain fewer repeats and no insertion/deletion PRH loci, leading to the absence of apparent length variants. However since allelic PRH1 variants were reported, mechanisms other than unequal pairing were proposed to explain the inter-individual PRH variations. Indeed, the third exons of the PRH1 and PRH2 genes were shown to contain only five tandem repeats but of substantial lengths differences. Such length differences were proposed to affect adversely the stability of misaligned DNA strands and to account for the production of PRH variants.

In addition, Maeda et al. [35] demonstrated that differential mRNA splicing and post-translational cleavages could generate a large number of PRPs that could explain the diversity of PR proteins. In the particular case of Peptide P-C Variant 1, one can hypothesize that the presence of this variant could derive from DNA slippage or differential mRNA splicing leading to the substitution of the GRPQGPPQQG-GH*QQGPPP* sequence for the GRPQGPPQQGGH-*PRPPR* sequence. This phenomenon can be reduced, due to the presence of a high sequence homology between tandem repeats, to the replacement of the QQGPPP amino acid string from the regular sequence toward PRPPR sequence (0.936 Da mass difference). Similarly, SNPs were also reported in salivary proteins with the exchange of Asp/Asn revealed by cDNA sequencing. The origin of the heavier form of Peptide P-C (Variant 2) could be explained by the occurrence of a SNP rather than from a deamidation event.

The biological role of these variants is obscure. We identified the three forms of Peptide P-C in several individuals (five samples from different volunteers). Apart from the role of PRPs in oral calcium disposition, relatively little is known about Peptide P-C from a biological perspective. Recent studies demonstrated that it can modulate blood glucose levels after feeding by inducing insulin release and lowering glucagon levels [15, 16]. Additional studies are needed to elucidate the biological significance of each Peptide P-C variant.

Conclusion

In this study, we demonstrated that unprejudiced sample analysis using MS and MS/MS techniques or other approaches is required to discover new protein isoforms. Results we obtained also raised the question of biological compound diversity that could be much more extended than previously expected. Another important issue addressed in this paper is related to the following interrogation: "Are the experiments, methodologies, instrument capabilities (sensitivity, resolving power, dynamic range . . .), and data analyses procedures used sufficient to fully characterize a selected sample?" This point is of particular importance for the search of new biomarkers in pathologies or to obtain an exact view of protein diversity and function [36].

Acknowledgments

The authors gratefully acknowledge financial support from the NIH-NIDCR (U01 DE016275 to D.T.W. and J.A.L.) and the NIH/NCRR High-End Instrumentation Program (S10 RR023045 to J.A.L.).

Appendix A Supplementary Material

Supplementary material associated with this article may be found in the online version at doi:10.1016/j.jasms.2010.01.026.

References

1. Teixeira, E. H.; Napimoga, M. H.; Carneiro, V. A.; de Oliveira, T. M.; Cunha, R. M.; Havt, A.; Martins, J. L.; Pinto, V. P.; Goncalves, R. B.; Cavada, B. S. In Vitro Inhibition of *Streptococci* Binding to Enamel Acquired Pellicle by Plant Lectins. *J. Appl. Microbiol.* **2006,** *101,* 111–116.
2. Hardt, M.; Thomas, L. R.; Dixon, S. E.; Newport, G.; Agabian, N.; Prakobphol, A.; Hall, S. C.; Witkowska, H. E.; Fisher, S. J. Toward Defining the Human Parotid Gland Salivary Proteome and Peptidome: Identification and Characterization Using 2D SDS-PAGE, Ultrafiltration, HPLC, and Mass Spectrometry. *Biochemistry.* **2005,** *44,* 2885–2899.
3. Messana, I.; Cabras, T.; Inzitari, R.; Lupi, A.; Zuppi, C.; Olmi, C.; Fadda, M. B.; Cordaro, M.; Giardina, B.; Castagnola, M. Characterization of the Human Salivary Basic Proline-Rich Protein Complex by a Proteomic Approach. *J. Proteome Res.* **2004,** *3,* 792–800.
4. Gomez, S. M.; Nishio, J. N.; Faull, K. F.; Whitelegge, J. P. The Chloroplast Grana Proteome Defined by Intact Mass Measurements from Liquid Chromatography Mass Spectrometry. *Mol. Cell. Proteom.* **2002,** *1,* 46–59.
5. Forbes, A. J.; Patrie, S. M.; Taylor, G. K.; Kim, Y. B.; Jiang, L.; Kelleher, N. L. Targeted Analysis and Discovery of Post-Translational Modifications in Proteins from Methanogenic Archaea by Top-Down MS. *Proc. Natl. Acad. Sci. U.S.A.* **2004,** *101,* 2678–2683.
6. Wu, S.; Lourette, N. M.; Tolic, N.; Zhao, R.; Robinson, E. W.; Tolmachev, A. V.; Smith, R. D.; Pasa-Tolic, L. An Integrated Top-Down and Bottom-Up Strategy for Broadly Characterizing Protein Isoforms and Modifications *J. Proteome Res.* **2009,** *8,* 1347–1357.
7. Messana, I.; Loffredo, F.; Inzitari, R.; Cabras, T.; Giardina, B.; Onnis, G.; Piludu, M.; Castagnola, M. The Coupling of RP-HPLC and ESI-MS in the Study of Small Peptides and Proteins Secreted In Vitro by Human Salivary Glands that are Soluble in Acidic Solution. *Eur. J. Morphol.* **2003,** *41,* 103–106.
8. Whitelegge, J. P.; Zabrouskov, V.; Halgand, F.; Souda, P.; Bassilian, S.; Yan, W.; Wolinsky, L.; Loo, J. A.; Wong, D. T.; Faull, K. F. Protein-Sequence Polymorphisms and Post-translational Modifications in Proteins from Human Saliva using Top-Down Fourier-transform Ion Cyclotron Resonance Mass Spectrometry, *Int. J. Mass Spectrom.* **2007,** *268,* 190–197.
9. Castagnola, M.; Congiu, D.; Denotti, G.; Di Nunzio, A.; Fadda, M. B.; Melis, S.; Messana, I.; Misiti, F.; Murtas, R.; Olianas, A.; Piras, V.; Pittau, A.; Puddu, G. Determination of the Human Salivary Peptides Histatins 1, 3, 5 and Statherin by High-Performance Liquid Chromatography and by Diode-Array Detection. *J. Chromatogr. B Biomed. Sci. Appl.* **2001,** *751,* 153–160.
10. Castagnola, M.; Inzitari, R.; Rossetti, D. V.; Olmi, C.; Cabras, T.; Piras, V.; Nicolussi, P.; Sanna, M. T.; Pellegrini, M.; Giardina, B.; Messana, I. A Cascade of 24 Histatins (Histatin 3 Fragments) in Human Saliva. Suggestions for a Pre-Secretory Sequential Cleavage Pathway. *J. Biol. Chem.* **2004,** *279,* 41436–41443.
11. Inzitari, R.; Cabras, T.; Onnis, G.; Olmi, C.; Mastinu, A.; Sanna, M. T.; Pellegrini, M. G.; Castagnola, M.; Messana, I. Different Isoforms and Post-Translational Modifications of Human Salivary Acidic Proline-Rich Proteins. *Proteomics* **2005,** *5,* 805–815.
12. Inzitari, R.; Cabras, T.; Rossetti, D. V.; Fanali, C.; Vitali, A.; Pellegrini, M.; Paludetti, G.; Manni, A.; Giardina, B.; Messana, I.; Castagnola, M. Detection in Human Saliva of Different Statherin and P-B Fragments and Derivatives. *Proteomics* **2006,** *6,* 6370–6379.

13. Fanali, C.; Inzitari, R.; Cabras, T.; Fiorita, A.; Scarano, E.; Patamia, M.; Retruzzelli, R.; Bennick, A.; Messana, I.; Castagnola, M. Mass Spectrometry Strategies Applied to the Characterization of Proline-Rich Peptides from Secretory Parotid Granules of Pig (*Sus scrofa*). *J. Sep. Sci.* **2008,** *31,* 516–522.
14. Messana, I.; Cabras, T.; Pisano, E.; Sanna, M. T.; Olianas, A.; Manconi, B.; Pellegrini, M.; Paludetti, G.; Scarano, E.; Fiorita, A.; Agostino, S.; Contucci, A. M.; Calo, L.; Picciotti, P. M.; Manni, A.; Bennick, A.; Vitali, A.; Fanali, C.; Inzitari, R.; Castagnola, M. Trafficking and Postsecretory Events Responsible for the Formation of Secreted Human Salivary Peptides: A Proteomics Approach. *Mol. Cell. Proteom.* **2008,** *7,* 911–926.
15. Kimura, I.; Sasamoto, H.; Sasamura, T.; Sugihara, Y.; Ohgaku, S.; Kobayashi, M. Reduction of Incretin-Like Salivatin in Saliva from Patients with Type 2 Diabetes and in Parotid Glands of Streptozotocin-Diabetic BALB/c Mice. *Diabetes Obes. Metab.* **2001,** *3,* 254–258.
16. Kimura, M.; Nakashima, N.; Kimura, I. Salivary Peptide P-C Modulates Both Insulin and Glucagon Release from Isolated Perfused Rat Pancreas. *Jpn. J. Pharmacol.* **1990,** *52,* 579–585.
17. Wolff, A.; Begleiter, A.; Moskona, D. A Novel System of Human Submandibular/Sublingual Saliva Collection. *J. Dent. Res.* **1997,** *76,* 1782–1786.
18. Whitelegge, J. P.; Zhang, H.; Aguilera, R.; Taylor, R. M.; Cramer, W. A. Full Subunit Coverage Liquid Chromatography Electrospray Ionization Mass Spectrometry (LCMS+) of an Oligomeric Membrane Protein: Cytochrome *b*(6)*f* Complex from Spinach and the Cyanobacterium *Mastigocladus laminosus*. *Mol. Cell. Proteom.* **2002,** *1,* 816–827.
19. Whitelegge, J. P.; Gundersen, C. B.; Faull, K. F. Electrospray-Ionization Mass Spectrometry of Intact Intrinsic Membrane Proteins. *Protein Sci.* **1998,** *7,* 1423–1430.
20. Roepstorff, P.; Fohlman, J. Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides. *Biomed. Mass Spectrom.* **1984,** *11,* 601.
21. LeDuc, R. D.; Taylor, G. K.; Kim, Y. B.; Januszyk, T. E.; Bynum, L. H.; Sola, J. V.; Garavelli, J. S.; Kelleher, N. L. ProSight PTM: An Integrated Environment for Protein Identification and Characterization by Top-Down Mass Spectrometry. *Nucleic Acids Res.* **2004,** *32,* W340–345.
22. Azen, E. A. Genetic Protein Polymorphisms in Human Saliva: An Interpretive Review. *Biochem. Genet.* **1978,** *16,* 79–99.
23. Azen, E. A. A Frequent Mutation in the Acidic Proline-Rich Protein Gene, PRH2, Causing a Q147K Change Closely Adjacent to the Bacterial Binding Domain of the Cognate Salivary PRP (Pr1′) in Afro-Americans. Mutations in Brief no. 154. Online. *Hum. Mutat.* **1998,** *12,* 72.
24. Hay, D. I.; Bennick, A.; Schlesinger, D. H.; Minaguchi, K.; Madapallimattam, G.; Schluckebier, S. K. The Primary Structures of Six Human Salivary Acidic Proline-Rich Proteins (PRP-1, PRP-2, PRP-3, PRP-4, PIF-s, and PIF-f). *Biochem. J.* **1988,** *255,* 15–21.
25. Scotchler, J. W.; Robinson, A. B. Deamidation of Glutaminyl Residues: Dependence on pH, Temperature, and Ionic Strength. *Anal. Biochem.* **1974,** *59,* 319–322.
26. Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W. Top-Down Versus Bottom-Up Protein Characterization by Tandem High-Resolution Mass Spectrometry. *J. Am. Chem. Soc.* **1999**, *121*, 806–807.
27. Pesavento, J. J.; Kim, Y. B.; Taylor, G. K.; Kelleher, N. L. Shotgun Annotation of Histone Modifications: A New Approach for Streamlined Characterization of Proteins by Top Down Mass Spectrometry. *J. Am. Chem. Soc.* **2004,** *126,* 3386–3387.
28. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated De Novo Sequencing of Proteins by Tandem High-Resolution Mass Spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2000,** *97,* 10313–10317.
29. Kelleher, N. L. Top-Down Proteomics. *Anal. Chem.* **2004,** *76,* 197A–203A.
30. (a) Siuti, N.; Kelleher, N. L. Decoding Protein Modifications Using Top-Down Mass Spectrometry. *Nat Methods* **2007,** *4,* 817–821; (b) Xu, P.; Peng, J.; Characterization of Polyubiquitin Chain Structure by Middle-Down Mass Spectrometry. *Anal. Chem.* **2008,** *80,* 3438–3444; (c) Carvalho, P. C.; Xu, T.; Han, X.; Cociorva, D.; Barbosa, V. C.; Yates, J. R. III. YADA: A Tool for Taking the Most Out of High-Resolution Spectra. *Bioinformatics* **2009,** *25,* 2734–2736.
31. Oppenheim, F. G.; Hay, D. I.; Franzblau, C. Proline-Rich Proteins from Human Parotid Saliva. I. Isolation and Partial Characterization. *Biochemistry* **1971,** *10,* 4233–4238.
32. Azen, E. A.; Oppenheim, F. G. Genetic Polymorphism of Proline-Rich Human Salivary Proteins. *Science* **1973,** *180,* 1067–1069.
33. Karn, R. C. Steroid Binding by Mouse Salivary Proteins. *Biochem. Genet.* **1998,** *36,* 105–117.
34. Herrera, J. L.; Lyons, M. F. II; Johnson, L. F. Saliva: Its Role in Health and Disease. *J. Clin. Gastroenterol.* **1988,** *10,* 569–578.
35. Maeda, N.; Kim, H. S.; Azen, E. A.; Smithies, O. Differential RNA Splicing and Post-Translational Cleavages in the Human Salivary Proline-Rich Protein Gene System. *J. Biol. Chem.* **1985,** *260,* 11123–11130.
36. Castagnola, M.; Messana, I.; Inzitari, R.; Fanali, C.; Cabras, T.; Morelli, A.; Pecoraro, A. M.; Neri, G.; Torrioli, M. G.; Gurrieri, F. Hypo-Phosphorylation of Salivary Peptidome as a Clue to the Molecular Pathogenesis of Autism Spectrum Disorders. *J. Proteome Res.* **2008,** *7,* 5327–5332.